Duration and intonation in emotional speech

Vroomen, J.H.M.; Collier, R.; Mozziconacci, S.J.L.

Published in:
Eurospeech 1993

Publication date:
1993

DURATION AND INTONATION IN EMOTIONAL SPEECH

Jean Vroomen[1], René Collier, and Sylvie Mozziconacci[2]

Institute for Perception Research, Eindhoven
[1] also University of Tilburg, Tilburg
[2] also University of Amsterdam, Amsterdam

ABSTRACT

Three experiments investigated the role of duration and intonation in the expression of emotions in natural and synthetic speech. Two sentences of an actor portraying seven emotions (neutral, joy, boredom, anger, sadness, fear, indignation) were acoustically analyzed. By copying pitch and duration of the original utterances to a monotonous one, it could be shown that both factors were sufficient to express the various emotions. In the second part, rules about intonation and duration were derived and tested. These rules were applied to resynthesized natural speech and synthetic speech generated from LPC-coded diphones. The results showed that emotions can be expressed accurately by manipulating pitch and duration in a rule-based way.

Keywords: *emotion, intonation, duration, synthetic speech*

INTRODUCTION

Prosodic features such as pitch, temporal structure, loudness, and voice quality serve a communicative purpose. A speaker may indicate, through prosodic means, to which information the listener should pay particular attention (accentuation, emphasis), and he may provide cues about the syntactic organization of the utterance (phrasing). Intuitively, however, the communicative function of prosody is most readily associated with the expression of attitude and emotion. Yet, at present, not much is known about the detailed correspondence between prosodic features and affect.

Past research of emotive speech has paid most attention to the contribution of variations in pitch, duration, loudness, and voice quality as measured in natural or in simulated affective speech. There is general agreement that if prosodic features are ranked in terms of their contribution, gross changes in pitch do contribute most to the transmission of emotions, whereas loudness seems to be least important [1]. Despite this general agreement, however, a major weakness remains that one has not been able to determine for each emotion a unique set of defining prosodic characteristics.

An omission of previous research is that one has not been able to provide a description of the intonation patterns of different emotions. Although it is agreed that pitch plays an important role, most studies have only looked at rather global statistical measures of pitch, like mean, range, or variability. Substantial progress might be obtained, if it could be shown how different emotions can be expressed by distinct intonation patterns. One aim of the present study is to analyze the pitch contours of emotional utterances and to classify them into intonation patterns. Ultimately, we want to derive a set of duration and intonation rules for creating emotive speech.

The study consists of three parts. Firstly, a database is set up of various emotional utterances that are recognized well above chance level. It will be examined whether duration and intonation of these utterances are salient cues by copying the time structure and/or the pitch contours from the emotional speech samples onto a monotonous utterance. The monotonous utterance is expected to become emotional if the transplanted prosodic component is an effective cue. In the second part, rules about intonation and duration are formulated that can generate emotional utterances from neutral speech. Thus, instead of copying the exact time structure and intonation contour, they are generated by rules. Finally, these rules are applied to LPC-coded diphone speech.

EXPERIMENT 1: Are Intonation and Duration Sufficient to Express Emotions?

A data base was set up from which speech samples could be

selected that were uniformly recognized as being good examples of certain emotional expressions. Three professional actors were asked to read out eight semantically neutral sentences with thirteen emotions: neutral, joy, happiness, boredom, worry, anger, sadness, fear, guilt, disgust, haughty, indignation, and rage. Eighteen subjects listened to the utterances and chose among the thirteen emotional labels the one that they thought was expressed. From the best speaker, the two best sentences ('zij hebben een nieuwe auto gekocht'; 'THEY HAVE BOUGHT A NEW CAR', and 'zijn vriendin kwam met het vliegtuig'; 'HIS GIRLFRIEND CAME BY PLANE') with the emotions neutral, joy, boredom, anger, sadness, fear, and indignation were selected for further analyses. All samples were recognized correctly at or above 50%, which is well above chance level. In the first experiment, we investigated whether pitch and/or duration were sufficient to express the various emotions.

METHOD

The pitch contour and/or the time structure was copied from the emotional utterances (source) onto a monotonous one (target), resulting in three conditions: duration-only, pitch-only, and duration-plus-pitch interchanges. The algorithm to copy the time structure and the pitch contour of the emotional utterances is a sort of 'waveform vocoder' based on a time-domain PSOLA algorithm [2 and 3]. It uses Dynamic Time Warping (DTW) for the proper time alignment between the source and target utterances. Intonation and duration are manipulated by means of PSOLA.

Subjects. Eight subject chose, after hearing an utterance, the emotional label that they thought was appropriate.

RESULTS

Table 1 presents the mean proportion of correct responses, pooled across the two sentences. An ANOVA was performed on the proportion of correct responses with type of interchange (3) and emotion (7) as within-subjects factors. There was a significant effect of type of interchange [$F(2,14) = 90.5$, $p < .001$], as the proportion of correct responses was .31 for the time interchange, .51 for the intonation interchange, and .81 for the time-plus-intonation interchange.

The main effect of emotion was significant [$F(6,42) = 11.41$, $p < .001$], as was the interaction between interchange and emotion [$F(2,84) = 13.41$, $p < .001$]. Inspection of Table 1 suggests that copying only the time structure of the emotional utterances onto the monotonous one can be effective for neutral, boredom and anger. Copying only the time-warped intonation contour of the emotional utterances is effective for boredom, sadness, fear, and indignation. If both time and intonation contours are transplanted, performance rises far above chance level for all emotions. The latter result suggests that one can generate highly recognisable emotions (on average 81 percent correct) from monotonous natural speech, if time and intonation are correct. This offers a perspective for superimposing emotions on diphone or allophone-coded synthetic speech, because it is apparently possible to express various emotions solely by controlling time and pitch, without the necessity to change the spectral composition of the utterance.

The next step was in a more rule-based direction. It was investigated whether general recipes can be discovered by which a neutral utterance can be converted into an emotional one. For that purpose, time and intonation rules were formulated.

EXPERIMENT 2: Rule Based Intonation and Duration

Based on several acoustic and perceptual analyses, it appeared that all emotional utterances obeyed the grammar of Dutch intonation. Hence, it was possible to substitute the original pitch contour of the emotional utterances with a 'standard' contour generated by the rules of the Dutch grammar [4]. The original contour could be replaced by an artificial contour without seriously affecting the emotional content of the message. Three free parameters needed to be set: the type of the intonation pattern (e.g., flat hat), the excursion size, and the key in the register. The optimal duration of the utterances was determined by linear compression, using the PSOLA technique. In the present study we applied these rules to a neutral utterance.

METHOD

Stimuli. The neutral utterance was linearly compressed/-expanded by means of PSOLA according to previously determined optimal compression rates. Then, the vowel-onsets of the to be accented syllables were determined manually. The appropriate software computed a pitch contour, taking into account the intonation pattern, excursion size, and key.

Subjects. Ten subjects chose, after hearing an utterance, the emotional label that they thought was appropriate.

RESULTS

The mean proportions of correct responses, pooled across the two sentences, are presented in table 2. On average, the utterances were correctly identified 55 percent of the time. An ANOVA with emotion (7) as within-subjects variables showed that there were differences among the emotions [$F(6,54) = 11.45$, $p < .001$]). Inspection of the table shows that in particular sadness and fear were recognized worse than the other emotions, but these and all other emotions were still recognized well above chance level. The next experiment investigates whether this still holds when speech is synthesized with LPC-coded diphones.

EXPERIMENT 3: Generating Emotions from Diphone Speech

In the present experiment, we investigated whether synthetic diphone speech can be emotionally coloured by changing duration and intonation.

METHOD

The utterances were synthesized from a phonetic description, using the IPO text-to-speech system [5]. The phonemes were converted into LPC-coded diphones, and standard duration rules were applied to these diphones. The resulting string of diphones was considered to have a duration that is appropriate for the neutral emotion. This utterance was then, for each emotion, linearly compressed or expanded using the previously determined optimal compression rates. The location of the accented syllables was marked manually, and an appropriate intonation pattern for the utterance was computed using the optimal parameter settings for excursion size and key. Two diphone sets were available from two different male speakers. For speaker 1, the diphones were coded in LPC with 12 poles, for speaker 2 it were 18. The diphones were synthesized, and then the duration and intonation were manipulated. Twelve subjects participated.

RESULTS

The mean proportions of correct responses, pooled across the two sentences, are presented separately for each speaker in table 3. On average, 63 percent of the utterances was correctly identified. An ANOVA with emotion (7) and speaker (2) as within-subjects variables indicated that the utterances of speaker 2 were slightly better identified than those of speaker 1 (67 percent versus 60 percent, respectively), but this difference did not reach significance [$F(1,11) = 2.47$, $p = .14$]. There were differences among the emotions [$F(6,54) = 11.45$, $p < .001$], and this interacted with the speakers [$F(6,66) = 3.72$, $p < .005$]. Inspection of table 3 suggests that joy was better recognized in speaker 1, whereas neutral, anger, and sadness were better recognized in speaker 2. But the most important result is that our rudimentary prosodic recipes, applied to a conventional diphone concatenation scheme, produce highly recognisable emotions in synthetic speech.

DISCUSSION

This study focused on the role of prosody, in particular pitch and duration, for the expression of emotions. In the first experiment it was shown that these prosodic parameters are sufficient to express various emotions. Then it was determined whether intonation and duration could be standardized and expressed in rule-like algorithms. It turned out to be possible to translate the pitch patterns of the emotional utterances into highly stylized intonation contours which obeyed the intonation grammar of Dutch. In this grammar, one has the option to choose among an array of structurally distinct pitch contours in which one may vary the excursion size of the pitch movements and the register. These parameters were sufficient to control the intonation in an acceptable way. The second simplification was in the time structure of the emotional utterances. It turned out to be legitimate to neglect temporal variations at the segmental level, and to change only the gross overall speaking rate of a neutral utterance in a linear way.

The rules for intonation and duration were applied to natural and LPC-coded diphone speech. The general picture is that the various emotions were correctly recognized far above chance level. Emotions are thus signalled by an ensemble of prosodic variables which can be controlled in synthesized speech. The results are promising, especially if one realizes that emotions are usually also signalled by the context and the content of the spoken message. Text-to-speech systems in which emotions can be expressed have thus become possible.

REFERENCES

[1]: Frick, R. W. (1985). Communicating emotion: the role of prosodic features. *Psychological Bulletin, **97***, 412-429.

[2]: Verhelst, W., & Borger, M. W. M. (1991). Intraspeaker transplantation of speech characteristics: an application of waveform vocoding techniques and DTW. *Eurospeech '91*, 1319-1322.

[3]: Charpentier, F., and Moulines, E. (1989). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Eurospeech '89*, 2, 13-19.

[4]: Zelle, H. W., de Pijper, J. R., & 't Hart, J. (1984). Semi-automatic synthesis of intonation for Dutch and British English. *Proceedings of the Tenth International Congress of Phonetic Sciences, Utrecht, **IIB***, 247-251.

[5]: van Rijnsoever, P. (1988). A multilingual text-to-speech system. *IPO Annual Progress Report, **23***, 34-39.